Investigating the Relatedness of the Endangered Dogon Languages

Moran, Steven ; Prokić, Jelena

Abstract: In this article we apply up-to-date methods of quantitative language comparison, inspired by algorithms successfully applied in bioinformatics to decode DNA and determine the genetic relatedness of humans, to language data in an attempt to shed light on the current situation of a family of languages called Dogon, which are spoken in Mali, West Africa. Our aim is to determine the linguistic subgroupings of these languages, which we believe will shed light on their prehistory, highlight the linguistic diversity of these groups and which may ultimately inform studies on the cultural boundaries of these languages.

DOI: https://doi.org/10.1093/llc/fqt061

Literary and Linguistic Computing Advance Access published September 17, 2013

Investigating the relatedness of the endangered Dogon languages

Steven Moran
University of Marburg and University of Zurich

Jelena Prokić
University of Marburg

Abstract

In this article we apply up-to-date methods of quantitative language comparison, inspired by algorithms successfully applied in bioinformatics to decode DNA and determine the genetic relatedness of humans, to language data in an attempt to shed light on the current situation of a family of languages called Dogon, which are spoken in Mali, West Africa. Our aim is to determine the linguistic subgroupings of these languages, which we believe will shed light on their prehistory, highlight the linguistic diversity of these groups and which may ultimately inform studies on the cultural boundaries of these languages.

Correspondence:
Steven Moran, Seminar für Allgemeine Sprachwissenschaft, Plattenstrasse 54, Raum 206, University of Zurich, 8033 Zurich, Switzerland.
Email:
steven.moran@uzh.ch

1 Introduction

This year's theme of the Digital Humanities conference is 'digital diversity: cultures, languages and methods'. Our research fits in naturally with this motif because human cultures and languages are intimately intertwined with each other. What is less obvious is that culture and language both disseminate in the same ways: through genealogical descent and through areal contact. Although, it is incredibly difficult to track the dissemination of cultures among societies, there are tried and tested methods for determining the genealogical relatedness of languages.

In this article we apply up-to-date methods of quantitative language comparison, inspired by algorithms successfully applied in bioinformatics to decode DNA and determine the genetic relatedness of humans, to language data in an attempt to shed light on the current situation of a family of languages called Dogon, which are spoken in Mali, West Africa. Our aim is to determine the linguistic subgroupings of these languages, which we believe will shed light on their prehistory, highlight the linguistic diversity of these groups and which may ultimately inform studies on the cultural boundaries of these languages.

Some of the Dogon languages are spoken by relatively small groups and thus can be considered endangered languages. The study of endangered languages and elaborate cultures is particularly urgent at this time in history, so we start by describing in Section 2 factors of language endangerment and we provide an overview of Dogon. In Section 3 we describe the data that we use in this work and in Section 4 we discuss the theory of how languages can be compared to establish genealogical relatedness. In Section 5 we discuss modern methods of quantitative language comparison and how they have been adapted from biology to linguistics. In Section 6 we give the analyses and results of applying two quantitative methods, the normalized edit distance (NED) and LexStat approaches, to the Dogon data. In Section 7 we provide an evaluation of these approaches and discuss their shortcomings. Lastly, in Section 8 we give our conclusion.

doi:10.1093/llc/fqt061

2 Background

2.1 Language endangerment

Current estimates are that there are around 7,000 languages spoken in the world today (Lewis *et al.*, 2013). Since the early 1990s, linguists have estimated that half of the world's languages will no longer be spoken by the end of this century (cf. Hale *et al.*, 1992). Recent findings based on precise data from the Catalogue of Endangered Languages show that 43% of the world's languages are currently endangered and that the rate at which languages are disappearing, roughly —three to four per year, has highly accelerated in the last forty years (Campbell *et al.*, 2013).[1] These figures represent a *dramatic* loss of the number of languages and the range of linguistic diversity of humans. When a language is lost, humanity loses a piece of the human knowledge base, which includes an irreplaceable encoding of culture-specific ways of thinking (Harrison, 2007).

Most people are familiar with the idea of endangered species and the extinction of animals, that is, there are very few remaining animals, or none remaining at all, of a particular species due to factors such as the destruction of the species' habitat, over hunting or pollution. A lesser known fact is that languages are becoming extinct at a rapid rate. But why is it that languages become endangered and/or extinct?

There are several causes of language endangerment (cf. Austin and Sallabank, 2011), including natural catastrophes, famine, and disease. For example, the small number of speakers of languages of the Andaman Islands were seriously affected by the Indian Ocean earthquake tsunami in 2004. War and genocide are also a factor that can lead to language endangerment and extinction; for example the atrocities of genocide by colonists in Tasmania, where all indigenous Tasmanian languages have been lost. Another factor, repression and forced linguistic and cultural assimilation, has lead to a decrease in the number of Native American languages in both North and South America. There are also influences from culturally, politically, and economically dominant cultures and countries.

Often these causes of language endangerment overlap with each other, but they typically share common influences that lead to *language shift*. The degree of language shift in a community of speakers determines the intergenerational language transmission rate, which is one way of measuring language vitality and endangerment. For example, UNESCO has a language vitality and endangerment framework, given in Table 1.

Language endangerment is happening all over the world and it is of the utmost urgency that linguists document and describe these languages before they disappear.[2] In many places in West Africa, for example, there is language shift towards colonial languages like English and French or towards large lingua-francas like Arabic and Swahili. Basically, as an increasing number of people become economically mobile and educated, those people may try to provide better future opportunities to their children by teaching them the more 'powerful' (or prestigious) language as their mother tongue, instead of passing on their own smaller, lesser known language.[3] One area where this shift is occurring is in Dogon country, a region in Mali renowned for its secluded and geographically divided village communities, some of which are embedded in an escarpment, that is, a sandstone cliff that rises 500 m (∼1500 feet) above the plains below it.[4]

2.2 Dogon languages

Dogon languages are spoken predominantly in eastern Mali in West Africa. The Dogon people were made famous by Marcel Griaule, a French anthropologist who pioneered ethnography in France, and who worked with the Dogon between 1931 and 1956. Reportedly the Dogon had advanced astronomical knowledge of the Sirius binary star system, knowledge that is not possible without telescopes. Since then, the Dogon have been shrouded in controversy and mystery.

As late as 1989, Dogon appeared in reference books as if it were a single language, for example., Bendor-Samuel (1989). In 2004, an extensive sociolinguistic survey by Hochstetler *et al.* (2004) estimated no less than seventeen distinct languages and described the language family as highly internally divided. The standard encyclopedic reference on the world's languages, the Ethnologue,[5] recently increased the number of Dogon languages that it

Table 1 UNESCO's language vitality and endangerment framework

Degree of endangerment	Intergenerational language transmission
Extinct	There are no speakers left
Critically endangered	The youngest speakers are grandparents and older, and they speak the language partially and infrequently
Severely endangered	Language is spoken by grandparents and older generations; while the parent generation may understand it, they do not speak it to children or among themselves
Definitely endangered	Children no longer learn the language as L1
Vulnerable	Most children speak the language, but it may be restricted to certain domains (e.g. home)
Safe	Language is spoken by all generations; intergenerational transmission is uninterrupted

lists from 14 to 19 (Lewis, 2009; Lewis *et al.*, 2013), but this figure is still too low. Since 2004, much initial survey work on Dogon has been undertaken by Professor Jeffrey Heath's Dogon Languages Project (DLP),[6] which has lead to the 'discovery' of a web of divergent dialects, some of which have been raised to the status of distinct languages based on standard linguistic criteria. The team has identified between 80 and 100 locally named varieties that have been tentatively group into approximately 20–25 languages. Nevertheless, the current Dogon linguistic situation is not at all transparent and much work is needed to document and describe these languages. Dogon languages are very under-described, some are highly endangered, and all are genealogically not well established (Blench, 2005; Heath, 2008). And although, Dogon is generally considered to be a division of the vast Niger–Congo language family, much work remains to establish this firmly.

The DLP provides a tentative yet detailed inventory of known Dogon languages. There are currently twenty distinct languages grouped (crudely) into eight geographical regions, with no implications for genealogical subgrouping. The internal structure of the Dogon language family is unknown, as is the number of mutually unintelligible languages it contains. In fact, the Ethnologue gives a flat genealogical tree. The position of the Dogon languages relative to other African language families is also unclear because of Dogon's typological characteristics. Its lineage has long been disputed, as summarized in Table 2.[7]

Thus, the current Dogon linguistic situation is not at all clear. Table 3 provides an approximate number of Dogon speakers, based on recent fieldwork by members of the DLP.

Table 2 Historical classification of Dogon

Year	Classification (language family)	Author
1924	Nigéro-Sénégalais	Delafosse
1941	Voltaic (Eng. Gur)	Homburger
1948	Voltaic; Gurunsi	Baumann and Westermann
1951	Mandé	Holas
1952	Mandé	Delafosse
1952	Gur (Fr. Voltaic)	Westermann and Bryan
1953	Voltaic	Bertho
1953	Non-classified	de Tressan
1950/60	Gur	Calame-Griaule
1963	Gur	Greenberg
1971	Gur	Bender-Samuel
1981	Voltaic	Manessy
1981	Volta-Congo	Bendor-Samuel
1993	Unresolved; non-classified	Galtier
1994	Unresolved; non-classified	Plungian and Tembiné
2000	Ijo-Congo	Williamson and Blench
2009	Volta-Congo	Lewis

One important factor regarding the need to understand the linguistic situation in Dogon has to do with education. Over the decades there have been several initiatives by the Malian government that involved sociolinguistic studies to determine which particular Dogon should be the 'official' language for all Dogon people. Or in other words, the studies were aimed at determining which Dogon language was the most mutually intelligible by the most number of speakers of other Dogon languages. The 'official' Dogon language, currently determined to be Toro So by crude lexicostatical methods (Hochstetler *et al.*, 2004), is now the language in which all literacy and written materials for schools,

etc., are to be developed. However, as linguists know, determining how closely a pair of languages are related is no trivial task and forcing groups of different language speakers to learn the same language can lead to dire consequences for minority languages.

3 Dogon Languages Data

There is much language and cultural data about the Dogon people that was collected in the field and is made available online through the DLP website.[8] These data include photos of flora and fauna, photos of lexical senses that illustrate areas other than flora and fauna, and photos of villages and topographic features of Dogon country. The website also provides videos of festivals, events, music, food and drink preparation, agriculture, pottery, hunting and gathering, and much more. The project has an even stronger focus on disseminating language data via the website. Grammars, dictionaries, and typological discussions are made openly available, as is the raw lexical data in an Excel spreadsheet, an example of which is shown in Table 4.

The Dogon comparative spreadsheet, which contains nearly 9,000 rows (meanings) and 20 columns (languages), aligns the words across different Dogon languages by their meanings. As the example in Table 4 shows, there is an English concept, such as 'sheep' or 'grass, herbs', and in each cell the particular Dogon word(s) are given in phonetic transcription. These data are collected by the DLP team in the field from speakers of these languages through common elicitation methods practiced by field linguists. The transcriptions of these words follow principles of the International Phonetic Alphabet.

Even without linguistic training and without knowing exactly what sound each letter stands for, the astute reader can glean some relations among the different words in the different languages in Table 4. For example the word 'sheep' in Jamsay and Togo Kan are the same, and in Togo Kan and Tomo Kan, it looks like their words for 'sheep' are not so different. On the other hand, compare the seemingly many different words in these different languages for 'grass' or 'herbs'. This is the type of reasoning that is behind historical and comparative linguistics.

Table 3 Approximate number of Dogon speakers by language

Tomo Kan	132,800	Tebul Ure	3000
Jamsay (?)	130,000	Toro Tegu	2900
Togo Kan (4)	127,000	Walo (2)	2000
Toro So (?)	50,000	Yanda Dom	2000
Donno So	45,300	Bangeri Me[a]	1200
Tommo So	40,000	Ampari Pa	1000
Najamba-Kindige (?)	24,700	Bunoge	500
Mombo	24,000	Ana	<500
Dogul Dom	12,200	Nanga	Unknown
Ambaleenge	6000	Tiranige	Unknown
Nyambeenge	5000	Pena	Unknown

Language names that are followed by a number or a question mark in parentheses indicate that there are one or more distinct languages under the same language name.

[a]Bangeri Me is a language isolate (along the lines of Basque) that is spoken in Dogon country, but that does not seem to belong to the Dogon language family. It may be an even more ancient language that no longer has any surviving relative languages.

4 Language Comparison and the Historical-comparative Method

Similar to species in biology, languages also evolve. Words are lost, new words are gained, and the pronunciation of all words changes slightly from day to day. During its history, a language may split into two or more descendant languages when the speakers separate and their languages keep on changing independently. To uncover how languages have

Table 4 Example of Dogon comparative wordlist

English	Toro Tegu	Nanga	Jamsay	Tommo So	Togo Kan	Tomo Kan
sheep	bélú	pèrgɛ́	péjú	pédú	péjú	pèjí
grass, herbs	sòlɔ́	bèrí, sàwâ	ɔ̀yɔ́	kèrú, bèlú	gú-gúrú	gwìní

evolved into their current shape is one of the major tasks of historical linguistics.

Uncovering language history is an incredibly difficult task for many reasons. Foremost is the fact that language, specifically the speech stream, is a changing and variable acoustic and articulatory signal. In fact, many acoustic and articulatory phoneticians believe that one cannot characterize speech sounds with discrete and invariant symbolic representations (such as letters to represent sounds). Nevertheless, researchers need some type of tangible object, even if a necessarily abstract one, to undertake any type of comparison and analysis of languages. Consider the fact that no two persons' pronunciation is identical, nor at a very fine-grained level does anyone say the same sound in exactly the same way twice in his or her lifetime. It is quite obvious, however, that Spanish and Italian are more closely related than Spanish and Mandarin Chinese. One might ask then, at what level then can we compare languages?

When a language is documented and described for the first time (most languages do not have writing systems), someone has to identify the different sounds in the language. That someone, perhaps a linguist or a trained missionary, determines the specific sounds in that language which trigger contrasts in the ear of the listeners. For example, in English the sounds /l/ and /r/ contrast so that we recognize a tangible difference between the words 'lake' and 'rake'. However, these sounds do not contrast in every language, for example, Korean speakers do not readily distinguish between them.

What we know about uncovering language history is that it is very difficult. For example, there are only a few languages whose history is directly reflected in written sources. For the majority of the 7,000 or so languages spoken today, we would not know anything about their past if we did not have methods to infer their history. In order to uncover language history, the languages spoken today are typically manually searched for traces of common origin. Finding these traces, however, is an extremely complicated task.

Constructing historical scenarios involves comparing words from different languages and identifying cognates. Cognates are words from different languages that go back to a common ancestor word (compare German *Hund* 'dog' and English *hound*). Cognate words exhibit a specific kind of similarity which does not necessarily show up in the form of surface resemblances of the sounds that the words are made of, but rather in structural similarities of cognate words. This kind of similarity is also not easy to detect: German *Zahn* 'tooth' and English *tooth*, for example, are cognate, while Greek *mati* and Malay *mata* are not.

The ultimate goal of historical linguistics is to construct historical scenarios. How such a construction (or reconstruction) is usually carried out can be easily illustrated by comparing the following four words in German, English, Italian, and French, which all mean 'tooth'.[9] Each is presented in Figure 1 in the same phonetic alphabet. In the first step, the words are all compared with each other, and common sound segments are analyzed, such as /d/ to /d/ and /t/ to /ts/.

In a second step, the sound sequences are 'aligned', that is, they are arranged in such a way that all corresponding segments occur in the same column, as illustrated in Figure 2.

Based on these identified correspondences, proto-stages of the languages are reconstructed. An ancestor word for both the German and English and the Italian and French word pairs is selected by certain principles that basically follow Occam's razor. This is shown in Figure 3.

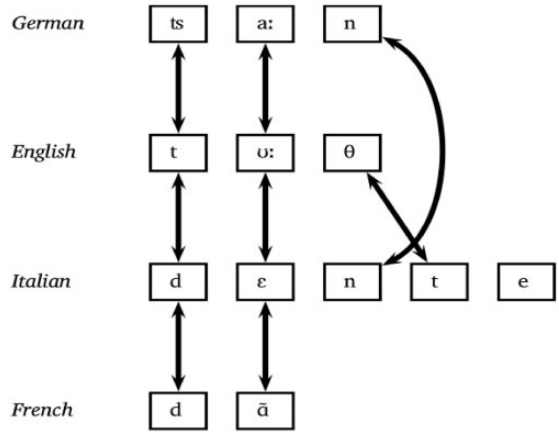

Fig. 1 Identify sound correspondences

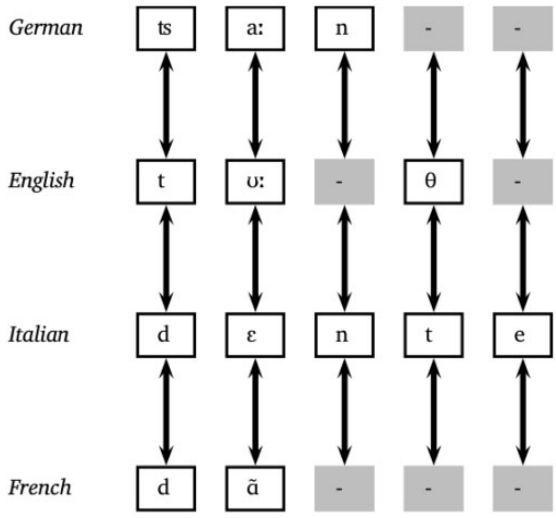

Fig. 2 Align the sounds

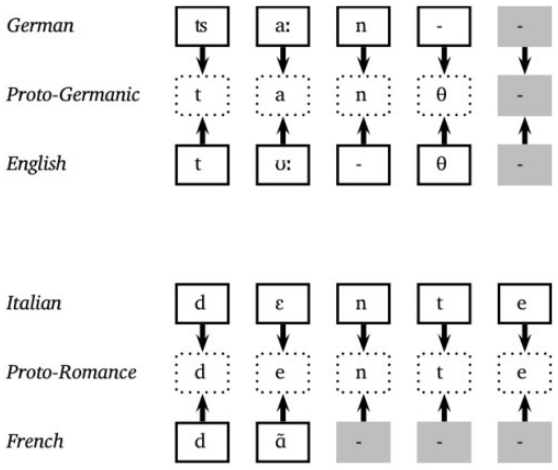

Fig. 3 Reconstruction of language proto-forms

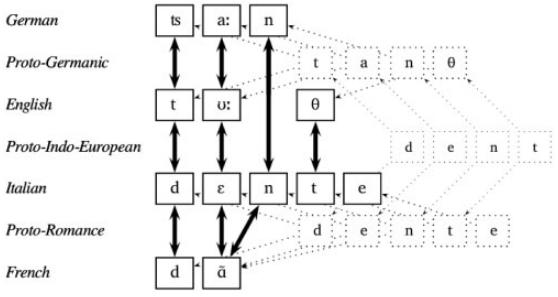

Fig. 4 Reconstructing a historical scenario

These reconstructed proto-stages can then further be compared. Finally, a historical scenario which shows how the words evolved into their current shape is reconstructed, as illustrated in Figure 4.

Historical reconstruction along these lines has typically been done by manual inspection by scholars with deep knowledge of different languages and their words' etymologies. Today, scientists are increasingly using quantitative approaches to automate and speed up this laborious and time-consuming task. Some of these quantitative approaches have been co-opted from biology, where inspiration has been taken from comparing sequences of DNA.

5 Quantitative Language Comparison

In both biology and historical linguistics, there are several parallels. First, sequences are defined as ordered collections drawn from a fixed set of characters; in molecular biology sequences constitute the basic unit of replication, in languages sequences constitute words. Second, whether DNA strings or words in different languages, sequence comparison is of paramount importance because it is the main method for determining relatedness between two objects. Lastly, both evolutionary biology and historical linguistics use phylogenetic trees (family trees) as their basic classification scheme. The former describe the evolution of species, and the latter, the evolution of languages.

There are several problems with these parallels between evolutionary biology and historical linguistics. Although the comparison problem is similar, the amino acid alphabet for proteins has only 20 characters. Compare this with the languages of the world, which have more than 2,000 different sounds (Moran, 2012). Additionally, biological sequences (such as proteins) are very long, while linguistic sequences (words) are relatively very short in comparison. And lastly and most importantly, in biology the alphabet remains stable during evolution, while it changes constantly in language history. Thus algorithms in biology were designed for long

sequences drawn from small alphabets. On the other hand, in quantitative historical linguistics we need algorithms for short sequences drawn from large alphabets.[10] To address these differences in biological comparison and language comparison, several different approaches are taken by linguists. In this article we will use two approaches to attempt to decode the relatedness of the Dogon languages: the normalized edit distance (NED) method and the LexStat method.

5.1 Normalized edit distance

Edit distance, also known as Levenshtein distance, is a metric used to measure the distances between two strings (Levenshtein, 1965). It represents the smallest number of edit operations (insertions, deletions, or substitutions) needed to transform one string into the other. At the same time, it aligns the two strings, as illustrated in Figure 5, which presents the alignment of two words for 'liver' taken from the Tebul Ure and Yorno So languages (two Dogon languages from our data). These two words differ in position 2 and position 4, where [ɛ́] corresponds to [í], and [d] has no corresponding sound (i.e. it corresponds to insertion or deletion of a sound), which means that the aggregate distance between these two strings is 2. In order to discard the influence of the lengths of the strings being compared, we normalize edit distance by dividing it by the length of the longer string.

The Levenshtein method has been extensively used in dialectometry to measure the distances between various dialects (Kessler, 1995; Nerbonne *et al.*, 1996; Heeringa, 2004). It has also been used to analyze the relatedness between languages, such as Indo-European (Serva and Petroni, 2008; Blanchard *et al.*, 2010), Austronesian (Petroni and Serva, 2008), and a very large sample of 3,002 languages (Holman, 2010). While dialect comparison based on the edit distance can give a pretty accurate picture of the aggregate distances between the dialectal varieties, some recent studies have shown the limits of this method when applied to languages at larger phylogenetic distances. For example, Prokić and Moran (forthcoming) compare three methods (the Levenshtein algorithm, n-gram approaches, and a very simple zipping technique) used to measure the distances between languages and show that these algorithms are not suitable for revealing deep genealogical relations on a set of sixty-nine indigenous South American languages. And Greenhill (2011) has shown that the accuracy of the Levenshtein method in classification of the Austronesian languages reaches only up to 65%. Furthermore, he has observed that the accuracy of Levenshtein classification decreases rapidly with phylogenetic distance.

```
k  é  n  d  é
k  í  n  -  é
-----------------
   1     1
```

Fig. 5 Illustration of two aligned strings

In this article, we rely on the NED method to shed more light on the synchronic relatedness among Dogon languages, without any reference to deep genealogical relatedness. We compare words of the same meaning by calculating their edit distance, and identify those groups of words that show 70% and higher similarity. We refer to the words that have the same meaning and a very similar form as 'homologies'. They could be real cognates or the result of internal or external borrowing. However, for the synchronic comparison of languages this is not relevant. To estimate the relation between each two language varieties in the data set, we calculate the number of shared 'homologies' between each two languages. The higher the number of shared homologies, the more similar two language varieties are and the higher mutual intelligibility between them. Since the current Dogon linguistic situation is not at all clear, we find this an important first step in comparing languages at the synchronic level, that is, for comparing the differences in languages as they are currently spoken. In order to automatically detect cognates and estimate genealogical relatedness between Dogon languages, we apply the LexStat approach described in the next section.

5.2 LexStat method

In recent years there have been several approaches to automatic cognate detection (Covington, 1996; Kondrak, 2002; Steiner *et al.*, 2011; Wettig *et al.*,

2012). In this research we rely on the LexStat method proposed by List (forthcoming, 2012b).[11] In the LexStat method each word is represented as a tuple of sound classes and prosodic strings. Sound classes are used to guess initial correspondences of sounds. The main idea of a sound class approach is that sounds which often occur in correspondence relations in genealogically related languages can be clustered into classes (or types). It is assumed 'that phonetic correspondences inside a ''type'' are more regular than those between different ''types''' (Dolgopolsky 1986: 35). Figure 6 provides an illustration of how a linguist might group sounds into classes.[12] Given a list of sounds, illustrated by /p, b, t, d, . . . /, sets of sounds can be grouped together into classes based on their regular sound correspondences between languages (List, forthcoming, chap. 4).

This process helps to reduce the number of sounds into a manageable size of twenty-eight classes (List, forthcoming, chap. 4.2.1), which is comparable to the number of characters used in biological algorithms. At the same time, for each word, a sonority profile is calculated. The sonority profile is used to derive a prosodic string, which is a vector of integer weights representing the relative sonority of segments in a sonority hierarchy (List, forthcoming, chap. 4.2.2). The prosodic strings combined with a scoring function helps account for the well-known fact that certain types of sound changes are more likely to occur in specific prosodic contexts. There are seven different prosodic environments distinguished in the LexStat method and each sound is assigned one of the seven values. In the next step, a permutation test is used to derive language-specific similarity scores between the sounds, and consequently between the words. This method compares the observed and expected distributions of the aligned sounds. While the observed distributions are calculated from the aligned words that show high phonetic and semantic similarity, that is, they can be considered potential cognates, the expected distributions are calculated from the randomly aligned words regardless of their phonetic or semantic similarity.

In the following section we use the NED and LexStat methods to investigate the relatedness of

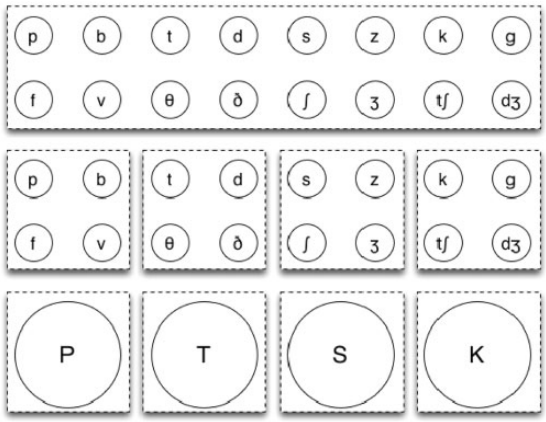

Fig. 6 Illustration of collapsing sounds into sound classes

the Dogon languages. The Dogon data are taken from the comparative lexical spreadsheet that is available online (as discussed in Section 3). From this spreadsheet we take a small subset of the data, based on the Swadesh word list, which includes words that are thought to be the most commonly found across languages, for example, 'sun', 'moon', 'man', and 'woman' (Swadesh, 1971).[13] We then parsed this list into a format that allowed us to run the NED and LexSat methods.

6 Analyses and Results

Our analysis is based on the 100 Swadesh list collected at nineteen Dogon villages and provided by the DLP. The distribution of the villages can be seen in Figure 7.

We apply the NED and LexStat approaches in order to align words and search for similar sets of words, both from the synchronic and diachronic perspectives. In both methods, once the similarities between the strings (words) are calculated, a certain threshold is required that determines which words are homologies/potential cognates. We determine this threshold for both approaches by empirically examining the aligned strings. The relatedness between languages is then determined by calculating the percentage of shared homologies/cognates. What we discovered is that while the LexStat

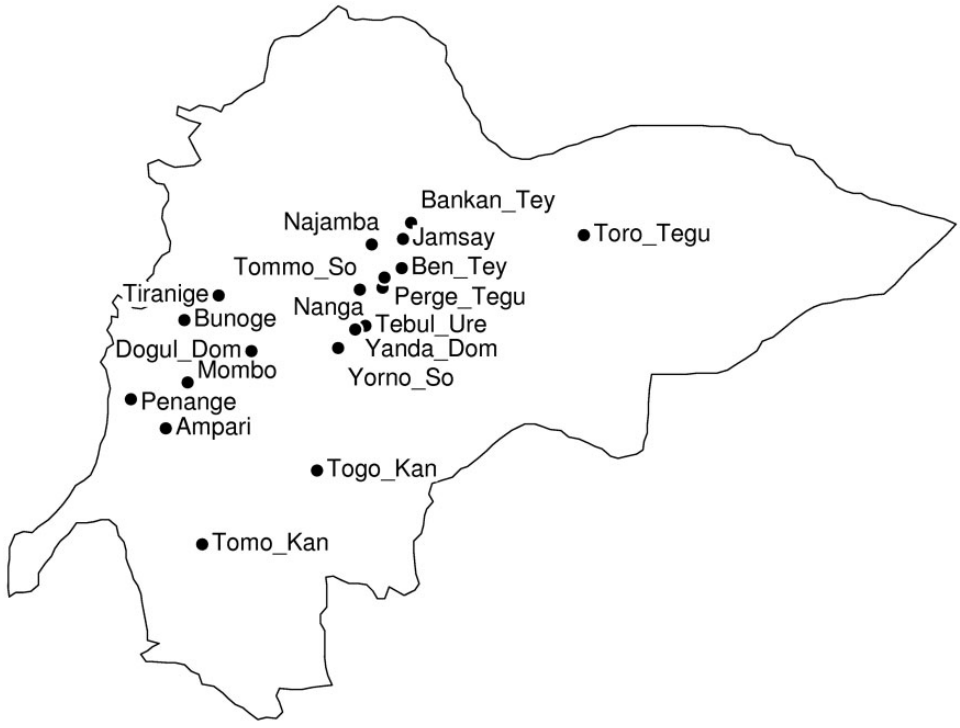

Fig. 7 Geographical distribution of villages where Dogon languages are spoken

method is designed specifically to discover cognates, the two approaches give almost identical results. In Table 5 we show the homologies identified by the NED method and the cognates identified by the LexStat method by inserting space between each group of identified homologies/cognates. The two columns list the lexical realizations of the same concept according to the NED method, on the left, and the LexStat method on the right.

The similarity matrices that express the percentage of shared homologies/cognates between each pair of languages are separately calculated using the NED and LexStat approaches. Each is first analyzed by means of *noisy clustering* (Nerbonne *et al.*, 2008). In *noisy* or *composite clustering,* small amounts of random noise are added to the matrices during repeated clustering in order to obtain stable clustering results.[14] The results of noisy clustering are probability dendrograms that show grouping of the data, as well as the confidence of the obtained groups (Figures 8 and 9).

With very high confidence, both analyses show separation of the western villages at the highest level (the biggest split in the data), where Mombo, Penange, Bunoge, Ampari, and Tiranige are spoken. The difference between the two methods is that in the LexStat method, unlike in the NED method, the village where Dogul Dom is spoken is not classified together with the other western villages. In order to see the geographical spread of the identified groups, we annotate with different shades the areas around each village in the data as determined by GPS coordinates for these villages. The areas around each village are determined by using Voronoi triangulation. The two way classification obtained by the two methods can be seen in Figures 10 and 11, where we use two different shades to show the division, dark gray for western and light gray for the eastern villages.

In the east, two groups of language clusters are identified by both methods (Figures 12 and 13). The difference is that in the NED analysis, the area

where Tomo Kan is spoken (9) is not classified in any of the two eastern clusters and we leave it shaded the same gray as the outlying area (Figure 12).[15] In the LexStat method, the area where Dogul Dom is spoken is not classified in any of the two eastern groups (also shaded the same as the outlying area in Figure 13).

The maps presented in Figures 12 and 13 reveal that the geographic spread of the identified language families is almost identical regardless of the method used. In the west, a homogeneous group of villages is clearly identified on both maps. The other two groups seem to be geographically mixed.

Table 5 Homologies identified by the NED method (left) and cognates identified by the LexStat method (right)

NED	LexStat
1. kûl	1. kûl
1. kwéː	
	2. kwéː
2. pìré	
2. pìré	3. bèrăː
	3. bèré
3. dólè	3. bèrɔ̀
	3. bèré
4. bèrăː	3. bèré
4. bèdè	3. pìré
4. bèré	3. bèré
4. bèrɔ̀	
4. bèré	4. dólè
4. bèré	
4. bèré	5. bèdè
4. pídʒí	5. pìndì
	5. bìndì
5. pìndì	5. bèté
5. bìndì	5. bèndé
6. bìn	6. pídʒí
6. bèté	
6. bèndé	7. dʒàŋgà
7. dʒàŋgà	8. bĭn

7 Evaluation

We evaluated classifications produced by our two approaches against a language family tree from the MultiTree project, based on Heath (2012) which represents genealogical classification of the Dogon languages.[16] We discarded the languages not present in our data set in order to make the comparison easier (Figure 14).[17]

The main division of the languages in Figure 14 separates the western and eastern groups. The western group includes five languages identified by our two methods, but also languages Tebul Ure, Najamba, and Yanda Dom, classified as eastern languages by both of our methods. The tree in Figure 14 is not well resolved and the only two subgroups of languages present are the Bunoge, Mombo, Ampari, and Penange cluster in the west, and Nanga, Ben Tey, and Bankan Tey in the east. These two smaller groups of languages can also be identified in the probability clusters in Figures 8 and 9. Geographic

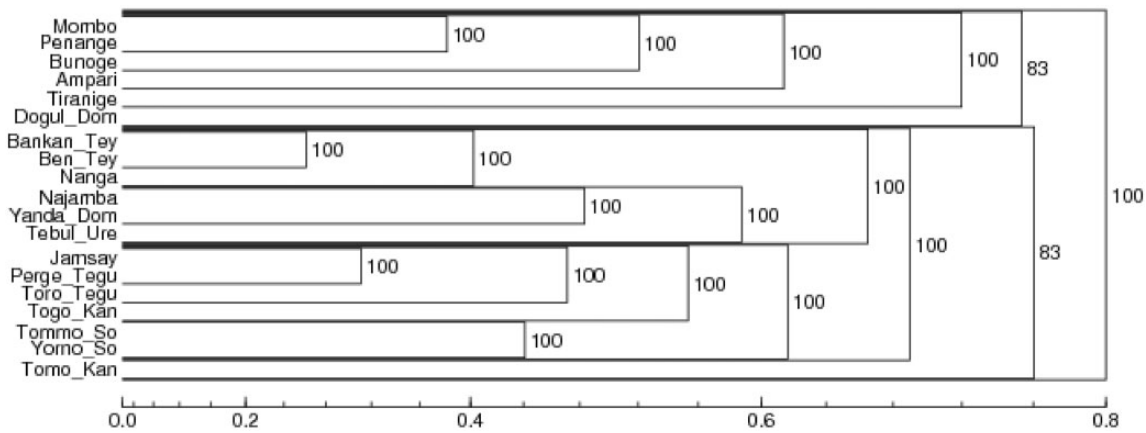

Fig. 8 Probabilistic dendrogram based on the NED approach

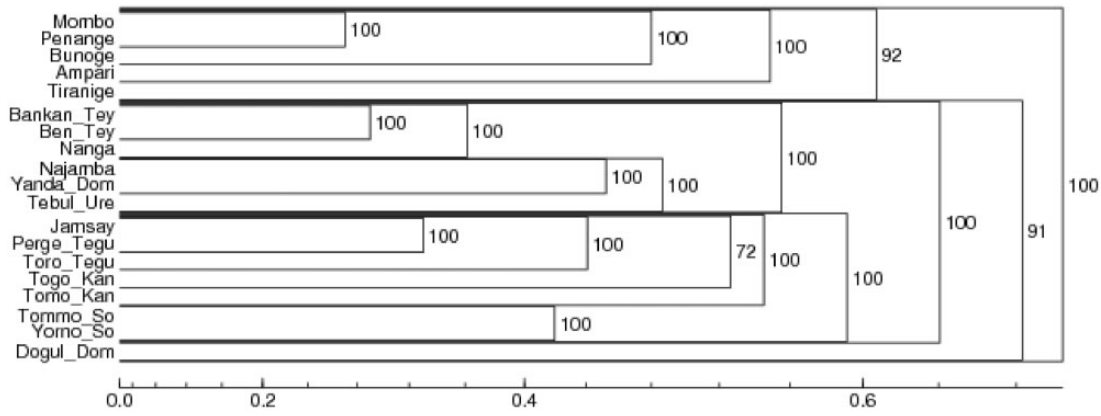

Fig. 9 Probabilistic dendrogram based on the LexStat approach

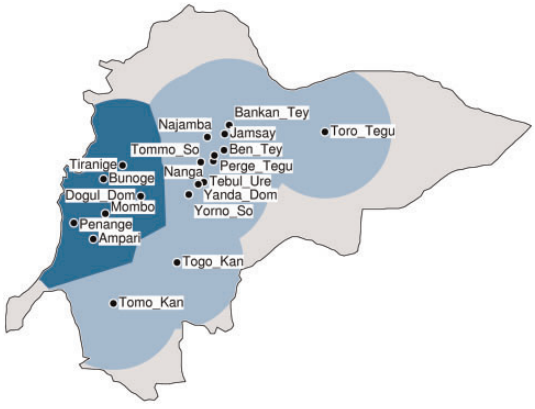

Fig. 10 Two groups of Dogon-speaking areas based on NED approach

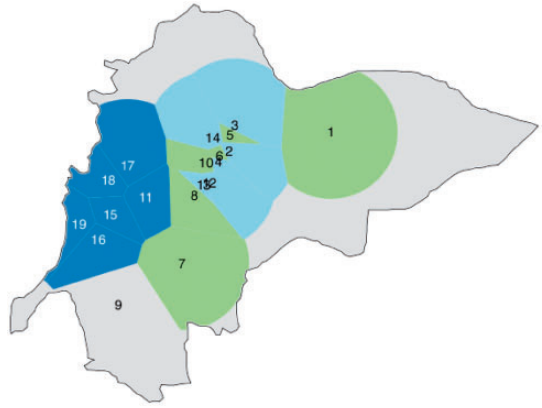

Fig. 12 Three groups of Dogon-speaking areas based on NED approach

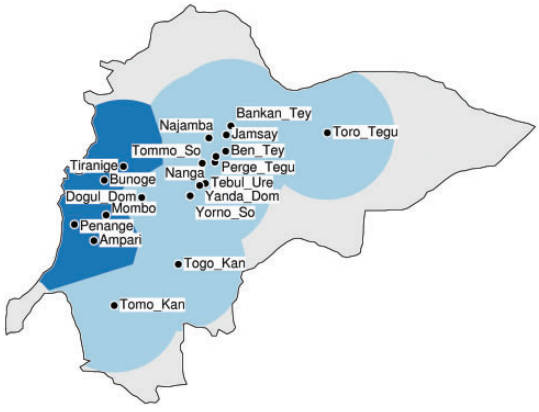

Fig. 11 Two groups of Dogon-speaking areas based on LexStat approach

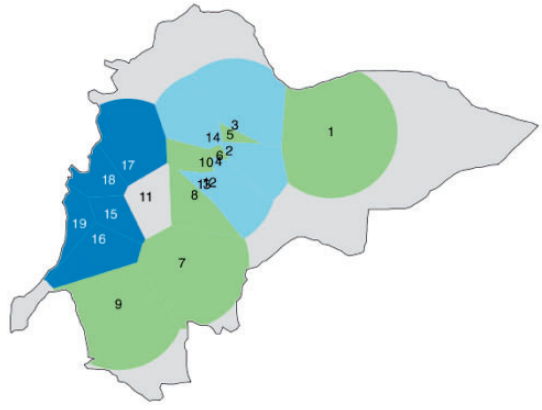

Fig. 13 Three groups of Dogon-speaking areas based on LexStat approach

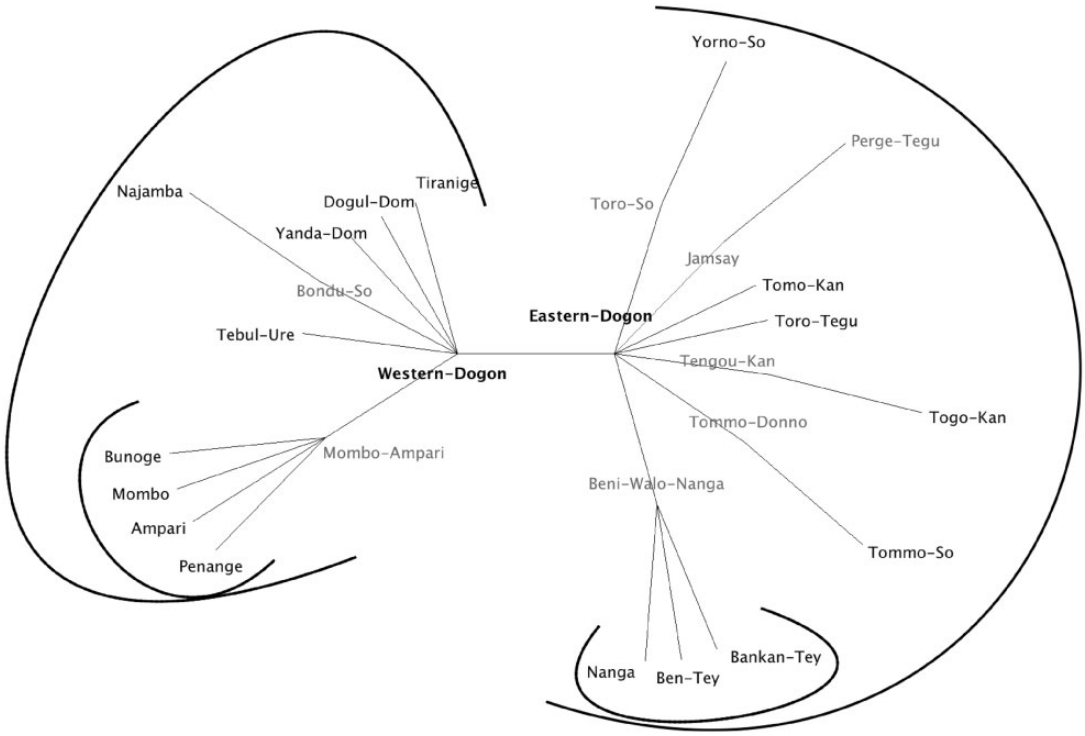

Fig. 14 Classification of languages based on Heath (2012)

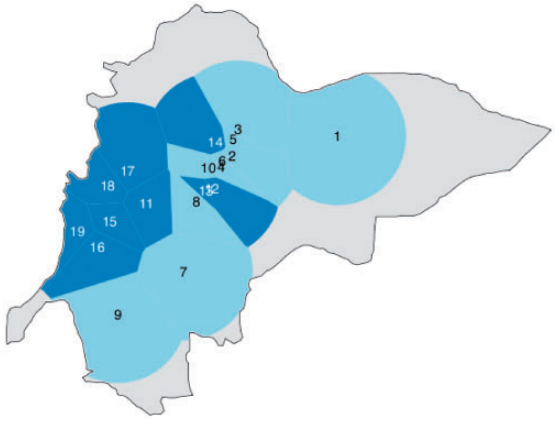

Fig. 15 Two-way classification of languages based on Heath (2012)

distribution of the main west-east division proposed by Heath (2012) can be seen in Figure 15.

The classifications suggested by the automatic methods offer a somewhat different view on the relatedness of the Dogon languages when compared to the genealogical classification. The difference is not only in the languages classified in the two main groups (western and eastern), but also in the finer subgroupings which are not present in the tree in Figure 14. The reason for some of the differences may be due to language contact, and subsequently language shift. As a result of physical movement of the whole villages to the east, Tebul Ure and Yanda Dom, genetically western languages, are nowadays deeply embedded in the area where eastern Dogon languages are spoken. Our analyses reveal that, probably due to contact with neighbouring languages, these genetically western varieties are shifting towards neighbouring languages at the lexical level. Najamba seems to be closely related to both Tebul Ure and Yanda Dom and lexically rather divergent from the immediately adjoining Dogon languages such as Jamsay and Tommo So. Our two automatic methods gave inconsistent classifications of Dogul Dom.

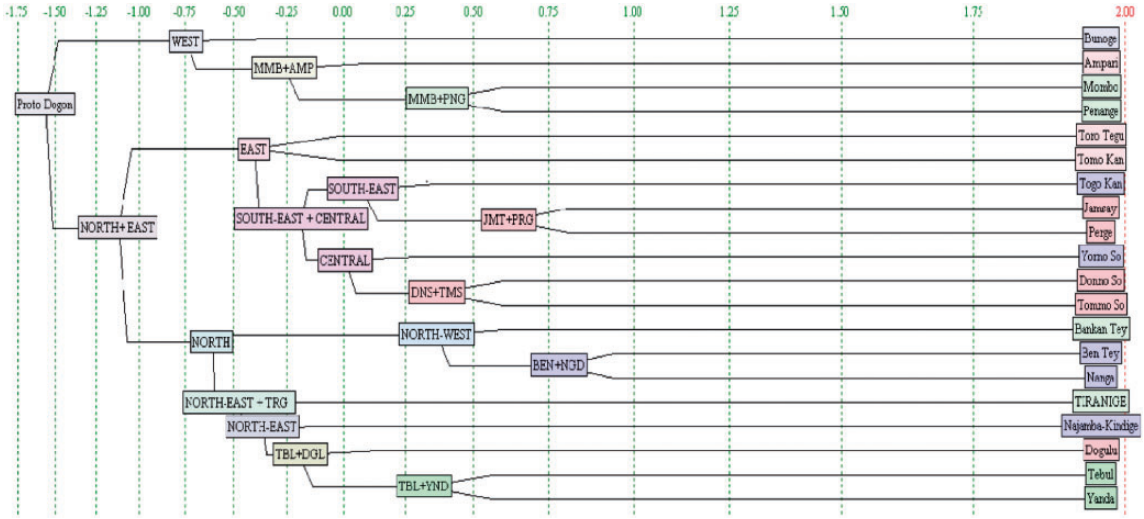

Fig. 16 Prokhorov's classification of Dogon languages

We also evaluated the classifications produced by the two automatic methods against a language family tree that was created by Prokhorov, an expert on Dogon languages (Prokhorov *et al.*, 2012). These trees are based on the number of shared 'lookalikes', that is, the set of true cognates and words that look alike due to either the borrowing of words between languages or pure chance. The lookalikes were manually selected by Prokhorov and his genealogical tree of Dogon is given in Figure 16.

Both automatic classifications that we used show an almost identical grouping of languages as the analysis done by Prokhorov. Two main groups of languages are clearly identified: western and eastern that is further divided into eastern and northern. Both computational methods gave almost identical results when compared to the opinion of an expert, with the difference that our method is fully automated and it can easily handle large amounts of data.

The results of both automatic methods suggest that (1) the major classification of Dogon languages follows the east-west division and that (2) Yanda Dom, Tebul Ure, and Najamba are lexically closely related to the neighboring eastern varieties probably due to language shift. Groupings obtained by the NED and LexStat methods nicely correspond to the classification suggested by Prokhorov and support it even at the level of dialect variation. Considering that research on the Dogon languages is still in its early stages, both when it comes to the traditional scholarship and quantitative methods, we find our work is an important first step in resolving the relatedness of the Dogon languages from both synchronic and diachronic perspectives.

8 Conclusion

In this article we have given a brief overview of language endangerment and the Dogon languages. Our goal has been to use quantitative methods repurposed from biology and applied to linguistics to automate methods used in traditional historical linguistics to compare and classify the relatedness of languages. Our goal fits well within digital humanities scholarship; we use digital methods to study and uncover knowledge about languages, and perhaps ultimately, their speakers' cultures. Our goal has been to try and untangle the mystery of how the different Dogon languages are related and we have identified some of the limitations of the current quantitative methods for language comparison.

These quantitative methods can only be seen as the first step in producing an automatic analysis

towards solving the genealogical relatedness of languages in general. Closer inspection of the potential cognates identified by the LexStat method reveals that the method is successful in cases where the two words that are compared show similar forms. However, this method alone cannot deal with the potential problem of borrowed words and additional analyses are needed. While for the synchronic comparison of languages this is not an important issue, for in-depth reconstruction of the historical relatedness of languages, we need to refine these digital methods even further in order to detect borrowings. However, the results of the two tested automatic methods are encouraging and we hope that they will be an important contribution in solving the problem of the diversity of Dogon languages.

Software

NED and LexStat analyses were performed using the LingPy software.[18] We use the GabMap software to produce Figures 7-13 and 15 (Nerbonne *et al.*, 2011).[19]

Funding

This work was supported by the ERC starting grant 240816: 'Quantitative modeling of historical-comparative linguistics'.

References

Austin, P. K. and Sallabank, J. (eds), (2011). *The Cambridge Handbook of Endangered Languages.* Cambridge University Press.

Bendor-Samuel, J., Olsen, E. J., and White, A. R. (1989). Dogon. In Bendor-Samuel, J. (ed.), *The Niger-Congo Languages: A Classification and Description of Africa's Largest Language Family*. Lanham, Maryland: University Press of America, pp. 169–77.

Blanchard, P., Petroni, F., Serva, M., and Volchenkov, D. (2010). Geometric Representations of Language Taxonomies. *Computer Speech and Language*, **25**(3): 679–99.

Blench, R. (2005). A Survey of Dogon Languages in Mali: Overview. *OGMIOS*, **26**: 14–15.

Campbell, L. (1998). *Historical Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

Campbell, L. *et al.* (2013). New Knowledge: Findings from the Catalogue of Endangered Languages. Paper presented at the 3rd International Conference on Language Documentation and Conservation (ICLDC). University of Hawai'i at Mānoa. February to March 2013.

Covington, M. A. (1996). An Algorithm to Align Words for Historical Comparison. *Computational Linguistics*, **22**: 481–96.

Greenhill, S. (2011). Levenshtein Distances Fail to Identify Language Relationships Accurately. *Computational Linguistics*, **37**(4): 689–98.

Hale, K. *et al.* (1992). Endangered Languages: On Endangered Languages and the Safeguarding of Diversity. *Language*, **68**(1): 1–42.

Harrison, K. D. (2007). *When Languages Die: The Extinction of the World's Languages and the Erosion of Human Knowledge*. New York and London: Oxford University Press.

Heath, J. (2008). *A Grammar of Jamsay*. Berlin: Mouton de Gruyter.

Heath, J. (2012). *Dogon and Bangime linguistics.* http://dogonlanguages.org/ (accessed 1 June 2013).

Heeringa, W. (2004). *Measuring Dialect Pronunciation Differences Using Levenshtein Distance.* Ph.D. thesis, University of Groningen.

Hochstetler, J. L., Durieux, J., and Durieux-Boon, E. (2004). *Sociolinguistic Survey of the Dogon Language Area.* SIL Electronic Survey Reports 2004, SIL International.

Holman, W. E. (2010). Do Languages Originate and Become Extinct at Constant Rates? *Diachronica*, **27**(2): 214–25.

Kessler, B. (1995). Computational Dialectology in Irish Gaelic. In *Proceedings of the European chapter of the Association for Computational Linguistics (EACL)*, pp. 60–6.

Kondrak, G. (2002). *Algorithms for Language Reconstruction.* Ph.D. thesis, University of Toronto.

Ladefoged, P. (1992). Another View of Endangered Languages. *Language*, **68**(4): 809–11.

Levenshtein, V. (1965). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Doklady Akademii Nauk SSSR*, **163**: 845–8.

Lewis, M. P. (2009). *Ethnologue: Languages of the World, Sixteenth Edition*. Dallas, TX: SIL International.

Lewis, M. P., Simons, G. F., and Fennig, C. D. (eds), (2013). *Ethnologue: Languages of the World, Seventeenth Edition*. Dallas, Texas: SIL International.

List, J.-M. (2011). Multiple Sequence Alignments in Historical Linguistics. Paper presented at the ConSOLE XIX: The 19th Conference of Student Organization of Linguistics in Europe. Groningen, Netherlands, January 2011.

List, J.-M. (2012a). Multiple Sequence Alignment in Historical Linguistics: A Sound Class Based Approach. In Boone, E., Linke, K., and Schulpen, M. (eds), *Proceedings of ConSOLE XIX 2011*. Leiden, The Netherlands: University of Leiden, pp. 241–60.

List, J.-M. (2012b). LexStat: Automatic Detection of Cognates in Multilingual Wordlists. In Butt, M., Prokić, J., Mayer, T., and Cysouw, M. (eds), *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL) 2012: Joint Workshop of LINGVIS & UNCLH*, pp. 117–25.

List, J.-M. (Forthcoming). *Sequence Comparison in Historical Linguistics*. Düsseldorf: Düsseldorf University Press.

List, J.-M. and Moran, S. (2013). An Open Source Toolkit for Quantitative Historical Linguistics. *Proceedings of the Association for Computational Linguistics (ACL) 2013*. Sofia, Bulgaria.

Moran, S. (2012). *Phonetics Information Base and Lexicon.* Ph.D. thesis, University of Washington.

Needleman, S. B. and Wunsch, C. D. (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology*, **48**(3): 443–53.

Nerbonne, J., Heeringa, W., van den Hout, E., van de Kooi, P., Otten, S., and van de Vis, W. (1996). Phonetic Distances between Dutch Dialects. In Durieux, G., Daelemans, W., and Gillis, S. (eds), *CLIN VI: Proceedings of the Sixth CLIN Meeting*. Antwerp: Centre for Dutch Language and Speech, pp. 185–202.

Nerbonne, J., Kleiweg, P., Heeringa, W., and Manni, F. (2008). Projecting Dialect Differences to Geography: Bootstrap Clustering vs. Noisy Clustering. In Preisach, C., Schmidt-Thieme, L., Burkhardt, H., and Decker, R. (eds), *Data Analysis, Machine Learning, and Applications*. Berlin, Heidelberg: Springer, pp. 647–54.

Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P., and Leinonen, T. (2011). Gabmap: A Web Application for Dialectology. *Dialectologia: revista electrònica*, **2**: 65–89.

Petroni, F. and Serva, M. (2008). Language Distance and Tree Reconstruction. *Journal of Statistical Mechanics: Theory and Experiment*, **2008**: P08012.

Prokhorov, K., Heath, J. and Moran, S. (2012). Dogon Classification. Paper Presented at the Niger-Congo Congress, Paris, September 2012.

Prokić, J. and Moran, S. (Forthcoming). Black Box Approaches to Genealogical Classification and Their Shortcomings. In Borin, L. and Saxena, A. (eds), *Approaches to Measuring Linguistic Differences*. Berlin, Boston: De Gruyter Mouton, pp. 429–45.

Serva, M. and Petroni, F. (2008). Indo-European Languages Tree by Levenshtein Distance. *Europhysics Letters*, **81**(6): 68005.

Steiner, L., Stadler, P. F., and Cysouw, M. (2011). A Pipeline for Computational Historical Linguistics. *Language Dynamics and Change*, **1**(1): 89–127.

Swadesh, M., Sherzer, J., and Hymes, D. (1971). *The Origin and Diversification of Language*. Chicago: Aldine De Gruyter.

Wettig, H., Reshetnikov, K., and Yangarber, R. (2012). Using Context and Phonetic Features in Models of Etymological Sound Change. In Butt, M., Prokić, J., Mayer, T., and Cysouw, M. (eds), *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL) 2012: Joint Workshop of LINGVIS & 75 UNCLH*, pp. 108–16.

Notes

1 http://www.endangeredlanguages.com/.

2 See the UNESCO Atlas of the World's Languages in Danger at: http://www.unesco.org/culture/languages-atlas/.

3 Ladefoged (1992) challenges the assumption that languages and cultures should be preserved, arguing that it is 'paternalistic of linguists to assume they know what is best' for a given speech community. Indeed, the adoption of major languages by certain small-language speech communities affords speakers better socio-economic opportunities. This issue is a fiercely debated in linguistics.

4 Dogon country has been recognized as a UNESCO World Heritage Site since 1989.

5 http://www.ethnologue.com/.

6 http://dogonlanguages.org/.

7 See Hochstetler *et al.* (2004) and Hantgan's Dogon bibliography for references at:http://dogonlanguages.org/bibliography.cfm.

8 http://dogonlanguages.org.
9 Figures 1–4 are taken, with permission, from List, forthcoming, and List 2012a.
10 Note also that while biologists can model their sequences in ASCII characters, linguists cannot easily do without the Unicode encoding due to the large number of different characters needed for phonetic transcription.
11 Here we give a high-level overview of the SCA procedure. A step-by-step tutorial with explanation is given with the Dogon data in the LingPy tutorial: http://lingpy.org/tutorial/. See also List & Moran, forthcoming.
12 Figure 6 is adapted, with permission, from List 2011.
13 An anonymous reviewer points out that Swadesh wordlists, which come in several versions that differ in length and semantic content, have been challenged as appropriate vehicles for lexicostatistics and glottochronology (cf. Campbell 1998). In general we agree with this criticism. For our study, we could have in principle used any set of words from the Dogon comparative wordlist because we know in general that Dogon languages are very closely related, both genealogically and geographically, and we have not attempted to do any dating of their genealogical relatedness. The Swadesh 100 list, as adapted slightly by the Dogon language experts to Dogon culture and geography (e.g. original 'earth' to 'earth (as brick mix or to repair walls)' or 'cold (of weather)' to 'cold (e.g. water)'), is nicely present in all villages in our survey.
14 In biology, in order to obtain stable clustering results a bootstrap procedure is often employed by randomly resampling the observed data (Felsenstein, 2004). Tested on dialect data, bootstrapping and noisy clustering produce distance matrices that correlate nearly perfectly ($r = 0.997$). Unlike bootstrapping, noisy clustering can be applied on a single distance matrix.
15 In Figures 12, 13, and 15 we replace language areas with numbers in order to make the grouping of eastern areas visible: Toro_Tegu (1), Ben_Tey (2), Bankan_Tey (3), Nanga (4), Jamsay (5), Perge_Tegu (6), Togo_Kan (7), Yorno_So (8), Tomo_Kan (9), Tommo_So (10), Dogul_Dom (11), Tebul_Ure (12), Yanda_Dom (13), Najamba (14), Mombo (15), Ampari (16), Tiranige (17), Bunoge (18) and Penange (19).
16 http://multitree.org/trees/Dogon%3A%20Heath%202012.
17 Classification of the Dogon language in the Ethnologue is flat and cannot be used to compare the groups obtained by out two methods.
18 http://lingpy.org/.
19 http://www.gabmap.nl/.